Assignment 3 - CSC/DSC 265/465 - Spring 2019 - Due March 28

Q1: For this question you will need to install the package `wooldridge` from the `CRAN` package repository (`https://cran.r-project.org/web/packages/`). Using default settings, this can be done using the command `install.packages("wooldridge")`. From this package we will use the data set `alcohol`:

```
> install.packages("wooldridge")
> library(wooldridge)
> data("alcohol")
```

(a) We will first make use of the following variables from the `wooldridge` data frame.

- `abuse`: =1 if abuse alcohol
- `mothalc`: =1 if mother an alcoholic
- `fathalc`: =1 if father an alcoholic

The binary response will be $Y = 1$ if the subject has a history of alcohol abuse (`abuse == 1`). Consider the logistic function $\phi(x) = (1 + \exp(-x))^{-1}$. Then set

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

where $X_1 = I\{\texttt{mothalc == 1}\}$, $X_2 = I\{\texttt{fathalc == 1}\}$. Then consider the logistic regression model:

$$P(Y = 1 \mid X_1, X_2) = \phi(\eta). \tag{1}$$

We will use the following notation for the odds ratio, given three events A, B, C.

$$OR(A \mid B, C) = \frac{odds(A \mid B)}{odds(A \mid C)} = \frac{P(A \mid B)/(1 - P(A \mid B))}{P(A \mid C)/(1 - P(A \mid C))}.$$

If C is omitted, we will assume $OR(A \mid B) = OR(A \mid B, B^c)$.

Show that for model (1) the odds ratio $OR(\texttt{abuse == 1} \mid \texttt{mothalc == 1})$ does not depend on the value of `fathalc`, and that

$$\begin{aligned} &OR(\texttt{abuse == 1} \mid \{\texttt{mothalc == 1 and fathalc == 1}\}, \{\texttt{mothalc == 0 and fathalc == 0}\}) \\ &= \quad OR(\texttt{abuse == 1} \mid \texttt{mothalc == 1}) \times OR(\texttt{abuse == 1} \mid \texttt{fathalc == 1}). \end{aligned} \tag{2}$$

(b) We next add an interaction term to model (1):

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2. \tag{3}$$

Then suppose that

$$OR(\texttt{abuse == 1} \mid \{\texttt{mothalc == 1 or fathalc == 1}\}, \{\texttt{mothalc == 0 and fathalc == 0}\}) = e^{0.63}.$$

This means that the increase in the odds of alcohol abuse is the same if only one or both parents are alcoholics. What would the values of β_1, β_2 and β_3 then be?

(c) Fit both models (1) and (3).

(i) Report the odds ratios $OR(\texttt{abuse == 1} \mid \texttt{mothalc == 1})$ and $OR(\texttt{abuse == 1} \mid \texttt{fathalc == 1})$ from model (1), with approximate 95% confidence intervals.

(ii) Is there evidence that $\beta_3 \neq 0$ in model (3)? What does that say regarding whether or not equation (2) holds?

Q2: According to Benford's Law (or the *first-digit law*), the frequency distribution of the leading digit in many sets of numerical data is not uniform. Rather, smaller digits tend to occur more frequently. For example, '1' occurs with a frequency of about $p_1 = 0.30$, and '5' or '9' occur with frequencies of about $p_5 = 0.071$ or $p_9 = 0.046$, respectively. The digit '0' is not represented.

This has been observed in accounting documents, and can therefore be used to detect fraud, the reasoning being that forged numbers tend to be randomly selected, and therefore each leading digit would appear with equal frequencies $\alpha = 1/9$.

Suppose a classifier is to be developed based on the observed frequencies $X = (n_1, n_5, n_9)$ of the leading digits that are '1', '5' or '9' in a given document. Let $n = n_1 + n_5 + n_9$. We wish to classify an accounting document as **(A)uthentic** or **(F)orged**. The relative frequencies of these three digits (assuming only these are observed in the data) are

$$\begin{aligned} P_A &= \left(\frac{0.30}{(0.30 + 0.071 + 0.046)}, \frac{0.071}{(0.30 + 0.071 + 0.046)}, \frac{0.046}{(0.30 + 0.071 + 0.046)} \right) \\ P_F &= (1/3, 1/3, 1/3)\,, \end{aligned}$$

for the two classes. Let π_A be the prior probability that a document is authentic.

(a) Bayes' Theorem gives the posterior odds as:

$$Odds(\text{Document is Authentic} \mid X) = LR \times Odds(\text{Document is Authentic}).$$

Given an expression for the posterior odds as a function of $P_A, P_F, \pi_A, n_1, n_5, n_9, n$.

(b) Show that a Bayesian classifier can be constructed which predicts that an accounting document is authentic if the following rule holds:

$$a \times n_1 + b \times n_5 + c \times n_9 + d \geq 0$$

where a, b, c, d are constants which depend on P_A, P_F, π_A, n. Give these constants as precisely as possible.

(c) Suppose the observed data is $X = (n_1, n_5, n_9) = (7, 5, 8)$, and the prior probability is set to $\pi_A = 1/2$. What is the posterior probability, given evidence X, that the document is forged?

Q3: This problem will make use of the data sets `Pima.tr` and `Pima.te` from the `MASS` library. These data sets are actually a random division of a single data set into test and training data sets. So restore the original data set:

```
> Pima.all = rbind(Pima.tr, Pima.te)
```

The complete data set should have 532 rows (one row per human subject) and 8 columns. Column 8 is the factor `type` with levels `Yes` (subject is diabetic according to WHO criteria) and `No` (otherwise). The first seven columns contain numerical variables which may be used to predict diabetes. See `help("Pima.tr")` for details. The object of this problem is to develop a classifier for response `type` based on the remaining seven features.

(a) For each of the seven quantitative features perform a two-sample Wilcoxon (Mann-Whitney) rank sum test in order determine whether or not the distribution of the feature differs between the diabetes positive and negative groups. Report the P-values. What does this suggest regarding the possibility of building an accurate classifier?

(b) Standardize the seven features by using a log-transformation (use the function `log`, which returns the natural logarithm by default). Add 1 to the variable `npreg` to avoid evaluating a logarithm of 0.

(c) A *confusion table* is a contingency table of the form:

	true diabetes -ve	true diabetes +ve
predicted diabetes -ve	n_{11}	n_{12}
predicted diabetes +ve	n_{21}	n_{22}

Any record used for testing the classifier is placed in exactly one of the four cells. Express classification error CE, sensitivity *sens* and specificity *spec* in terms of the elements $(n_{11}, n_{12}, n_{21}, n_{22})$ of the confusion table. Create an `R` function that inputs the confusion table, and outputs a single vector with elements $(CE, sens, spec)$.

(d) Build a classifier using each of the following methods.

- **Linear discriminant analysis (LDA):** Use the `lda` function with option `CV = TRUE`.
- **Quadratic discriminant analysis (QDA):** Use the `qda` function with option `CV = TRUE`.
- **KNN classifier (KNN):** Use the KNN function `knn.cv`. Allow the neighborhood size K to vary over $K = 1, 3, \ldots, 33, 35$. Select as K the value which minimizes CE.

(i) What form of cross-validation is used by the `lda`, `qda` and `knn.cv` functions?

(ii) Using all features, report for each method the summary statistics $(CE, sens, spec)$. Is there one single classifier which is optimal with respect to all summary statistics?

(e) The three methods above give a binary classification. Many binary classifiers can be constructed from a quantitative score W. The binary classification follows by selecting a threshold t, so that a positive classification is equivalent to $\{W \geq t\}$. The values of *sens* and *spec* vary as t is varied, and an appropriate balance between the two types of error can be selected.

(i) Logistic regression can be used in this way, with $W = \phi(\eta)$, or equivalently $W = \eta$, where $\eta = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$. Fit a logistic regression model which predicts $I\{$`type == Yes`$\}$ based on the remaining 7 features (do not include interactions).

(ii) Create an ROC curve for this model, that is, plot the values of $(1 - spec, sens)$ as t varies. **HINT:** You can use code from `CLASSIFICATION-B.R`.

(iii) Superimpose the values $(1 - spec, sens)$ obtained for the three classifiers from Part (d) onto the ROC curve. When classifiers give very different balances between sensitivity and specificity, how can a plot such as this one help determine which classifier makes the most efficient use of the data?

(iv) For LDA/QDA, the classifier can be quantified indirectly by varying the prior class probabilities, using the `prior` option to introduce user specified values. This has the same effect as varying the threshold t. Refit the LDA and QDA classifiers, varying the prior probability of diabetes +ve π_+ over the grid `seq(0,1,0.01)`. Again, use the `CV = TRUE` option. For each fit, store *sens* and *spec*. Plot again the ROC curve for the logistic regression model, then superimpose ROC curves for the LDA and QDA classifiers. Does either the LDA or QDA seem preferable? Note that the logistic regression classifier was not evaluated by cross-validation, and here serves the purpose of a reference ROC.

Q4: [Graduate Students Only] *Random forest* classifiers are discussed extensively in Chapter 8 of *An Introduction to Statistical Learning* (James et al). Here we will use a well known `R` implementation, accepting the default options, to repeat the analysis of Question 3.

(a) Make sure the package `randomForest` is installed and loaded. The classifier can be fit with the command

```
fit.rf = randomForest(type ~ ., data=Pima.all)
```

then class predictions in the form of quantitative *class probabilities* can be obtained with the command

```
pred.class = predict(fit.rf, type='prob')
```

(b) Plot again the ROC curve for the logistic regression model and LDA and QDA classifiers from Question 3 (e)-(iv). Superimpose on this an ROC curve for the random forest classifier. Does this form of classifier offer any advantage over LDA or QDA for this application?